ACOUSTIC EVENT LOCALIZATION USING A CROSSPOWER-SPECTRUM PHASE BASED TECHNIQUE

Maurizio Omologo *Piergiorgio Svaizer*
IRST-Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo di Trento (Italy)

ABSTRACT

Linear microphone arrays can be employed for acoustic event localization in a noisy environment using time delay estimation. Three techniques are investigated that allow delay estimation, namely Normalized Cross Correlation, LMS Adaptive Filters, Crosspower-Spectrum Phase: they are combined with a bidimensional representation, the Coherence Measure, in order to emphasize information that can be exploited for estimating position of both non-moving and moving acoustic sources. To compare the given techniques, different acoustic sources were considered, that generated events in different positions in space. Expressing performance in terms of accuracy of the wavefront direction angle, experiments showed that the Crosspower-Spectrum Phase based technique outperforms the other two. This technique provided very promising preliminary results also in terms of source position estimation.

1. INTRODUCTION

In the last decade, some research effort has been devoted to microphone array processing techniques [1], especially for teleconferencing and large room recording [2], but also for speech recognition [3].

Recently, the use of microphone array technology for acoustic surveillance purposes has been considered [1]: the objective of this activity is detection of acoustic events (e.g. explosions, screams, etc.) that can occur in a given environment, as well as localization of the acoustic source that generated them. This paper will investigate the problem of source localization, when a linear microphone array (consisting of four omnidirectional microphones) is used for acquisition of such events in a real noisy environment.

From a theoretical point of view, the signals acquired by each microphone can be assumed to be delayed replicas of the source signal plus noise: localizing the sound source is equivalent to estimating the time delays between the signals received. Once the delays are known the acoustic event direction can be derived using geometry.

The localization technique described in this work consists in determining the source position as crossing point between directions estimated beginning from signals acquired by microphone pairs. This method requires a very accurate time delay estimation but results the most effective, as described in [8], when the acquisition array consists of a few microphones.

Three different time delay estimation techniques (Normalized Cross Correlation (NCC), LMS Adaptive Filters (LMS), Crosspower-Spectrum Phase (CSP)) were compared using both real environment signals and synthetically delayed and distorted replicas. Information on relative delays is summarized and visualized using a meaningful representation called Coherence Measure.

Results in terms of angle as well as source position accuracy showed a definite superiority of the CSP-based technique.

2. GEOMETRICAL MODEL

This section introduces the general sound source model, for a two dimensional geometry and a linear microphone array consisting of M acoustic sensors. The geometry of the array is represented by the sensor positions $(p_0(x_0,y_0)$, ..., $p_{M-1}(x_{M-1},y_{M-1}))$. We assume that an acoustic source located in position (x_s,y_s) (see Figure 1) generates an acoustic event $r(t)$ that is acquired by microphones $0,.....,(M-1)$ as signals $s_0(t)$, ..., $s_{M-1}(t)$.

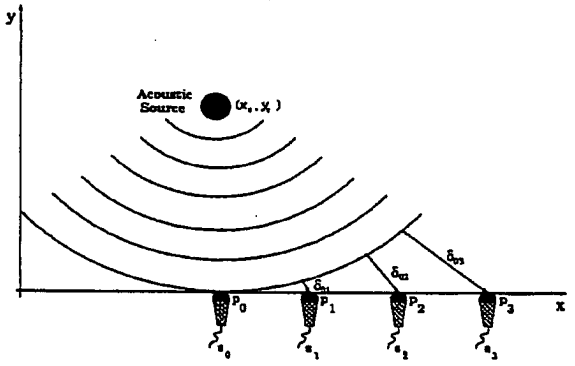

Figure 1. *Wavefront propagation of an acoustic stimulus generated in position* (x_s, y_s). *Signals* s_0, s_1, s_2, s_3 *are acquired through an array of microphones placed in positions* p_0, p_1, p_2, p_3. *The wavefront reaches microphones* 1, 2, 3 *with delays* δ_{01}, δ_{02}, δ_{03}, *with respect to microphone* 0.

For the given source signal $r(t)$, propagated in a generic noisy environment, the signal acquired by the acoustic sensor "i", can be expressed as follows:

$$s_i(t) = \alpha_i r(t - \tau_i) + n_i(t) \qquad (1)$$

where α_i is an attenuation factor due to propagation effects, τ_i is the propagation time and $n_i(t)$ includes all the contaminating noises, which are assumed to be uncorrelated with $r(t)$. We also indicate with $\delta_{ij}(x,y)$ the relative delay of wavefront arrival between microphones "i" and "j", assumed a source in position (x,y) and in particular:

$$\delta_{ij} = \delta_{ij}(x_s, y_s) = (\tau_j - \tau_i). \qquad (2)$$

[1] This work was partially supported by the ESPRIT 5345 DIMUS project, where a system is being developed for surveillance of underground stations.

3. COHERENCE MEASURE

As pointed out above, different techniques can be exploited for time delay estimation. Information on mutual delay between signals can be reconducted into a representation called Coherence Measure (CM) and associated to a function $C_{ij}(t,\tau)$ that expresses, given a delay τ the similarity between segments (centered at the time instant t) extracted from two generic signals s_i and s_j. It is expected to have a prominent peak at the delay $\tau = \delta_{ij}$, corresponding to the direction of wavefront arrival. For each couple of microphones, and for each technique, a bidimensional CM-representation can be conceived as shown in Figure 2: in this representation horizontal axis is referred to time, vertical axis is referred to delay and the coherence magnitude is represented through a grey scale. Following examples will show CMs evaluated with an analysis rate of 10.65 ms.

Both for moving and for non-moving sources, this CM can be exploited to derive the source position. When the acoustic source is moving, the CM maximum should depict a curve that follows the contour of the theoretical delay $\delta_{ij}(t)$ for the given microphone pair (i,j). For non-moving sources CM representation is characterized by a line at the theoretical delay; hence, the delay can be easily extracted.

4. ANALYSIS TECHNIQUES

Different methods can be conceived to derive the CM representation. In this section, three techniques are described, namely: Normalized Cross Correlation, LMS Adaptive Filters and Crosspower-Spectrum Phase.

4.1. Normalized Cross Correlation

The most common method of determining the generic time delay δ_{ij} and the corresponding arrival angle, given two signals $s_i(t)$ and $s_j(t)$, requires to estimate for every delay τ, the cross correlation function:

$$R_{ij}(\tau) = E[s_i(t)s_j(t+\tau)] \qquad (3)$$

where E denotes expectation. Given the model (1), (3) can be expressed as:

$$R_{ij}(\tau) = \alpha_i\alpha_j R_{\tau\tau}(\tau - \delta_{ij}) + R_{n_in_j}(\tau) \qquad (4)$$

where $R_{\tau\tau}(\tau)$ represents the autocorrelation of the source signal $\tau(t)$, evaluated at lag τ. δ_{ij} can be theoretically derived maximizing this function with respect to τ. However, due to the finite observation time, (3) can be only estimated for a given temporal window of length T, centered at time t; we denote this estimate as:

$$\widehat{R}_{ij}(t,\tau) = \frac{1}{T}\int_{t-T/2}^{t+T/2} s_i(u)s_j(u+\tau)du. \qquad (5)$$

When dealing with real noisy signals, the simple maximization based approach applied to (5) can easily fail, due both to signal properties (dependent also on microphone characteristics) and to limits of the mathematical model. A reasonable improvement [5] consists in the normalization of (5) with respect to the signal energies, leading to:

$$\widehat{R}^{(N)}_{ij}(t,\tau) = \frac{\int_{t-T/2}^{t+T/2} s_i(u)s_j(u+\tau)du}{\sqrt{\int_{t-T/2}^{t+T/2} s_i^2(u)du}\sqrt{\int_{t-T/2}^{t+T/2} s_j^2(u+\tau)du}}. \qquad (6)$$

In the following, we will refer to a first type of Coherence Measure defined as:

$$C'_{ij}(n,l) = \widehat{R}^{(N)}_{ij}(n,l). \qquad (7)$$

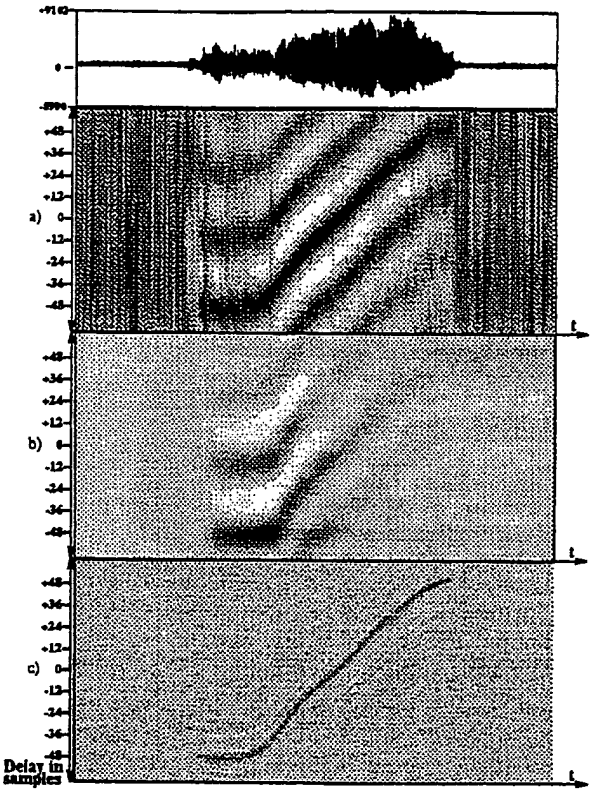

Figure 2. *Coherence Measure representation based on the NCC (a), the LMS (b) and the CSP (c) analyses between signals s_0 (plotted in the upper part of the Figure)and s_1: the acoustic stimulus was a vowel "a", uttered in a noisy environment by a speaker that was walking from the left to the right of the microphone pair.*

that indicates the digital counterpart of (6).

In an ideal situation, when (1) is simplified imposing both that $\alpha_i = 1$ for every i (i.e. there is no attenuation) and that noise components $n_i(t)$ are uncorrelated, $\widehat{R}^{(N)}_{ij}(n,l)$ has a peak equal to 1 for the lag corresponding to the delay δ_{ij}. But, if $\tau(t)$ is a periodic signal with period $T_P = PT_s$ (denoting sampling rate period with T_s), $\widehat{R}^{(N)}_{ij}(n,l)$ contains other unitary peaks, for each lag $\delta_{ij} + kT_P$ (with k integer): this property can be reconducted to the periodic characteristic of $R_{\tau\tau}(\tau)$. Also microperiodicities (e.g. due to formant structure of speech vowels) contribute to make evident other misleading peaks.

Further, [5] pointed out other problems that arise applying a simple peak-picking algorithm to CM expressed by (7): a critical issue is the window length T; in the following we will refer to the use of a window length of 21.3 ms (1024 samples at F_s= 48 kHz), that resulted a reasonable compromise between complexity and performance, especially for moving sources.

4.2. LMS Adaptive Filters

The LMS Adaptive Filter [6] is a Finite Impulse Response (FIR) filter that automatically adapts its coefficients to minimize the mean square difference between its two inputs. It does not require any a priori knowledge of the input spectra. The structure consists of two input signals: a reference

signal $s_j(t)$ and a desired signal $s_i(t)$; both signals can be modeled as (1)(their samples at time nT_s will be denoted with $s_i(n)$ and $s_j(n)$). The LMS Adaptive Filter output is based on the following formula:

$$y_{ij}(n) = W_{ij}^T(n)X_{ij}(n) \tag{8}$$

where T denotes transpose and $X_{ij}(n) = [s_j(n), s_j(n-1), ..., s_j(n-L+1)]^T$ is the filter state, consisting of the most recent samples of the reference signal. The vector $W_{ij}(n)$ is the L-vector of filter weights at instant n. The error output $\varepsilon_{ij}(n)$ is computed as follows:

$$\varepsilon_{ij}(n) = s_i(n) - W_{ij}^T(n)X_{ij}(n). \tag{9}$$

The weight vector is updated every sample:

$$W_{ij}(n+1) = W_{ij}(n) + \mu\varepsilon_{ij}(n)X_{ij}^*(n) \tag{10}$$

where $*$ denotes the complex conjugate and μ is a feedback coefficient that controls the rate of convergence and the algorithm stability.

The algorithm adapts the FIR filter to insert a delay equal and opposite to that existing between the two signals: in an ideal situation, the filter weight corresponding to the true delay would be unity and all other weights would be zero.

For our purposes and due to the properties of the weight vector $W_{ij}(n)$, we define a second type of CM as:

$$C''_{ij}(n,l) = w_{ij}(n,l) \tag{11}$$

where $w_{ij}(n,l_0)$ is the $W_{ij}(n)$ component for lag l_0.

4.3. Crosspower-Spectrum Phase

Starting from a mathematical modeling similar to (1), [7] proposed a maximum likelihood estimator for determining time delays between two signals s_i and s_k. Prefiltering signals before computing correlation, leads to the so-called Generalized Cross Correlation method. Basically, the Fourier transform of (4) provides the Crosspower-Spectrum:

$$G_{ik}(f) = \alpha_i\alpha_k G_{rr}(f)e^{-j2\pi f\delta_{ik}} + G_{n_in_k}(f). \tag{12}$$

The Generalized Cross Correlation between $s_i(t)$ and $s_k(t)$ is defined as:

$$R_{ik}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_g(f)G_{ik}(f)e^{j2\pi f\tau}df \tag{13}$$

where $\psi_g(f)$ is a general frequency weighting filter.

A way to sharpen the cross correlation peak is to "whiten" the input signals: the choice

$$\psi_p(f) = \frac{1}{|G_{ik}(f)|} \tag{14}$$

leads to the so-called phase correlation technique [7]. With such a choice:

$$R_{ik}^{(p)}(\tau) = \int_{-\infty}^{+\infty} \frac{G_{ik}(f)}{|G_{ik}(f)|}e^{j2\pi f\tau}df \tag{15}$$

and, if noise signals are uncorrelated it follows that:

$$\frac{G_{ik}(f)}{|G_{ik}(f)|} = e^{-j2\pi f\delta_{ik}}. \tag{16}$$

It is worth noting that, contrary to the other two delay estimate techniques, the generalized correlation given by (15) is independent from the input waveform characteristics, i.e. in the ideal case it reduces to a delta function centered at the correct delay δ_{ik}.

In practice, the procedure for estimating the generalized correlation starts from the computation of spectra $\widehat{S}_i(t,f)$ and $\widehat{S}_k(t,f)$ through Fourier transforms applied to windowed segments of s_i and s_k, centered around time instant t. Then, these power spectra are used to estimate the normalized Crosspower-Spectrum:

$$\phi(t,f) = \frac{\widehat{S}_i(t,f)\widehat{S}_k^*(t,f)}{|\widehat{S}_i(t,f)||\widehat{S}_k(t,f)|} \tag{17}$$

that preserves only information about phase differences between s_i and s_k. Finally, the inverse Fourier transform $\widetilde{R}_{ik}(t,\tau)$ of $\phi(t,f)$ is computed. Also in a real situation, the resulting function (defined in the lag axis τ) has a constant energy, mainly concentrated on the correct delay δ_{ik}. The Coherence Measure introduced in this case is $C'''_{ik}(n,l) = \widetilde{R}_{ik}(n,l)$.

4.4. Delay estimation algorithm

Given the CM representation, the source position can be derived in different ways: as pointed out, if the source is non-moving CM should consist of a dominant straight line at the theoretical delay. Hence, starting from Coherence Measure $C_{ij}(n,l)$ evaluated for microphone pair (i,j), for all delays l $(-l_{MAX} <= l <= l_{MAX})$ and time samples n $(1 <= n <= N)$, a lag can be estimated as follows:

$$\widehat{l}_{ij} = \arg\max_l[\sum_{n=1}^{N} C_{ij}(n,l)]. \tag{18}$$

Once given two delay hypotheses, obtained by processing two microphone pairs, source direction is estimated averaging angles corresponding to these delays, while source position is computed as crossing point between such directions.

5. EXPERIMENTS AND RESULTS

Different approaches can be followed to evaluate and compare performance that can be obtained with the three mentioned analysis techniques. As described in the following, two series of experiments were conducted, one operating on synthetic signals, the other on real signals, collected in a noisy environment. Both series confirmed the superiority of the CSP-based technique to the others.

5.1. Simulation Experiments

Given a real signal, originally acquired with a sampling frequency of 48 kHz from one of the array microphones, and another, obtained shifting the original one of a given delay, two new versions of these signals were artificially determined adding both different attenuated and shifted replicas (to simulate reverberation phenomena) and different white noise sequences, to better match the mathematical modeling introduced with (1).

During the simulation experiments, three real signals were considered, that is a syllable /pa/, a whistle sound and a long speech message. For each of them, 300 artificial signal couples were generated, using different combinations of attenuation factors, interchannel delays and white noise magnitude.

Using the CM-functions based on the three given techniques, delay estimates were computed by the previously described algorithm. Table 1 shows that the mean delay error and its standard deviation are very different, given the three techniques, even if they deal with artificial signals. In particular, using the NCC-based technique and, sometimes,

	Mean	StdDev
NCC	1.24	1.57
LMS	0.32	0.54
CSP	0.09	0.12

Table 1. *Mean and standard deviations of the lag error (expressed in samples) obtained applying the three techniques to 900 artificial signal couples.*

the LMS-based one, often provide a delay slightly different (one or two samples) from the theoretical one, that can cause an unacceptable performance in terms of wavefront arrival angle accuracy. This behavior can be explained as follows: these two techniques are strongly influenced by the presence of microperiodicities in the given signals; in this way, the corresponding CM-function curves depict a primary peak, that results from "interpolation" among peaks positioned both at the correct delay and at other signal-dependent spurious delays.

As previously mentioned, the CSP-based technique is free from this influence since it can be considered "independent" from the given signal characteristics.

5.2. Real Signal Experiments

In a first scenario a database of 97 stimuli was collected in an acquisition room, in some cases with a background noise previously recorded in an underground station to simulate a real-noise environment. In this case, acquisition was accomplished by using an array consisting of four equispaced microphones: distance between microphones was 15 cm. Approximately half of the stimuli were acquired in presence of background noise with an average SNR of 15 dB. The database consists of screams (the syllable /pa/ uttered loudly by one speaker), whistle-sounds, gun-shots. Acoustic events were generated in the following room positions (expressed in meters in x and y axes, respectively): (0,1), (0,2), (0,3), (1,1), (1,2), (1,3), (2,2), (2,3), (-1,1), (-1,2), (-1,3), (-2,0.5) (-2,1), (-2,2), (-2,3).

In terms of wavefront direction angle, performance confirms properties that were observed in previous discussions. In particular, given a reasonable tolerance of 5°, the CSP-based technique ensures performance 20% better than LMS and 40% better than NCC based techniques. However, performance in terms of localization accuracy were not considered satisfactory, especially using NCC and LMS based techniques. Table 2 shows performance expressed in terms of accuracy of the wavefront arrival angle, given three different tolerances of 2°, 5°, 10°.

	< 2°	< 5°	< 10°
NCC	35	48	56
LMS	59	67	70
CSP	66	88	96

Table 2. *Percentage of acoustic events that were correctly identified in terms of wavefront direction angle, given three different angle tolerances.*

To better investigate localization accuracy allowed by the CSP-based analysis technique, a second scenario was conceived using a new array configuration that consists of two distant microphone pairs: distance between microphones of each pair was 15 cm, while distance between microphone pairs was 75 cm.

A new database of 100 acoustic stimuli was acquired in an office environment. Acoustic events were generated in the following 20 room positions (expressed in meters in x and y axes, respectively): (-1,1), (-1,2), (-1,3), (-1,4), (0,1), (0,2), (0,3), (0,6), (0,7), (1,1), (1,2), (1,3), (1,4), (2,0.5), (2,1), (4,5), (-2,0.5), (-2,1), (-2,5), (-3,5). In each position we produced a gun shot, a handbeat, a scream, a short word and a whistle.

Performance in terms of wavefront direction angle were comparable with those obtained in the first scenario (90% accuracy given a tolerance of 5°). Concerning localization accuracy, given a tolerance of (15 cm, 50 cm), 63% of stimuli were correctly localized. Given a tolerance of (30 cm, 100 cm), 85% of stimuli were correctly localized.

These experiments confirm the good behaviour of the CSP-based technique but also the importance of the array geometry. Clearly, the use of more pairs is expected to provide further substantial improvement.

6. CONCLUSIONS

This work provided a comparison among three techniques of acoustic source localization. That based on Crosspower-Spectrum Phase has the best properties for the estimation of the wavefront arrival direction: it is the most robust both in clean and in non critical noisy environments. Issues that will be investigated are new microphone array configurations as well as robustness of this technique in more severe background noise conditions.

Even if in this work the computational aspect was not emphasized, the proposed technique offers many advantages also from this point of view: at the moment, the resulting localization system runs in real-time on a DSP-board equipped with two DSP32C and four acquisition channels, operating with a sampling frequency of 48 kHz and 16 bit accuracy.

Finally, the CSP based analysis is being investigated for other purposes as talker tracking and speech enhancement [8].

REFERENCES

[1] J. L. Flanagan, H. F. Silverman, "Material for International Workshop on Microphone Array Systems: Theory and Practice", Technical Report LEMS-113, Division of Engineering-Brown University, October 1992.

[2] J. L. Flanagan, J. D. Johnston, R. Zahn, G. W. Elko, "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms", J.Acoust.Soc.Am. 78(5), November 1985, pp.1508-1518.

[3] H. F. Silverman, "Some Analysis of Microphone Arrays for Speech Data Acquisition", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-35, n.12, December 1987.

[4] M. Omologo, P. Svaizer, "Use of the CrosspowerSpectrum Phase in Acoustic Event Localization", IRST Technical Report #9303-13, March 1993, submitted for publication.

[5] H. F. Silverman, S. E. Kirtman, "A Two-stage Algorithm for Determining Talker Location from Linear Microphone Array Data", Computer Speech and Language (1992) 6, pp.129-152.

[6] F. A. Reed, P. L. Feintuch, N. J. Bershad, " Time Delay Estimation Using the LMS Adaptive Filter-Static Behavior", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-29, n.3, June 1981.

[7] C. H. Knapp, G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-24, n. 4, August 1976.

[8] M. Omologo, P. Svaizer, "Talker Localization and Speech Enhancement in a Noisy Environment using a Microphone Array based Acquisition System", Proceedings Eurospeech'93, Berlin, September 1993.